PRAGMATIC PHONETICS: ACOUSTIC CORRELATES

Katherine Morton
University of Essex, Colchester, U.K.

ABSTRACT

Within the general theory of linguistics, pragmatics is concerned with describing the intentions, attitudes and beliefs of the speaker. Pragmatic phonetics itself is about how speech production interprets the requirement to communicate pragmatically determined effects, and about how the perceptual system is triggered by the acoustic signal to invoke the appropriate reaction in the listener. This paper examines the role of the acoustic signal in this complex chain of processes, and discusses how the system might usefully be modelled.

SPEECH PRODUCTION THEORY

Phonetics has been primarily concerned with modelling the physical processes of speech production. Speech production is usually associated with motor, aerodynamic and acoustic processes. But phonetics also models speech perception, involving physical and cognitive processes.

Modern theories of speech production blur the distinction between cognitive and physical processes [1] [2]. For example, they do this in different ways: Articulatory Phonology uses the gestural score formalism to represent requirements in both planning and execution: Cognitive Phonetics introduces cognitively driven supervision of motor processes to explain how some of the universal physical effects of speech vary in a linguistically sensitive way.

The rigid distinction between phonology and phonetics made it difficult to understand the effects of these phenomena. Integration of cognitive and physical descriptions at both the production and perception levels of speech modelling is essential if we want to examine pragmatic phenomena in speech.

PRAGMATIC PHONETICS

Pragmatics can be thought of as an extension of semantics [3]; linguistic semantics describes meaning and within semantics, pragmatics attempts to explain the interpretation of meaning in terms of attitudes and belief structures. Pragmatic phonetics models how the set of beliefs and intentions available to human beings become part of spoken language. It is concerned with the expression and interpretation (or production and perception) of intention, attitude and belief, where these properties of the language are not directly expressed by choice of words or word order in sentences, but by *how the utterance is said.*

This claim means that there must be different ways of speaking a particular sentence, and that the resulting acoustic signals will trigger in the listener an awareness of the emotion, attitude or belief which the speaker may be communicating [4].

Additionally this communication may not be voluntary, that is, under the conscious control of the speaker. For example, a speaker might be so angry as to be unable to suppress communicating that anger though tone of voice, or so happy that that emotion cannot be suppressed.

Pragmatic phonetics is therefore about triggering a listener response to stimuli over and above the usual phonological and phonetic content of utterances [5] [6].

TONE OF VOICE

Pragmatic Phonetics is being developed for two reasons:

- to characterize and explain pragmatically derived effects in speech production and perception, and
- to simulate the acoustic effects using speech synthesis.

It will be useful to have synthetic speech able to convey an added dimension of naturalness [7], but the simulation is also being developed to test the model itself. This rests on the assumption that it is possible to capture the tone of voice which triggers effects in the perceiver, and that the information is in the acoustic signal. Two consequences of this model are

- we can link semantics and phonetics;
- we can model the humanness of dialogue.

Perceived variations in tone of voice should be obvious and detectable as departures in the acoustic waveform from an expected norm. It should be possible in principle to identify and quantify these changes. But there is considerable variability in speech waveforms and it has proved difficult to separate out the variations associated with conveying pragmatic effects from other variability present in the waveform [8], introduced by properties of the vocal tract and articulators.

It is essential now to develop a computationally oriented model for synthesis. In modelling spoken communication for dialogue systems, it is useful to distinguish between two types of independently varying and independently sourced tone which produce different types of pragmatic effect. This choice enables the explicit execution in speech of pragmatic markers in speech. For a computational model to operate, this arrangement requires a hierarchical rather than linear organization.

1. Global tone of voice

Tone of voice at the global level characterizes what is appropriate for the overall dialogue situation. Here are some examples from human/machine dialogue situations:

- In an inquiry system about the weather the informant would ideally sound friendly and confident of the facts.
- In a situation warning of emergency the speaker needs to be simultaneously firm, confident and reassuring.
- In a dialogue as part of a computer assisted teaching programme, the listener should be made to feel that the speaker is being sympathetic as well as instructive.
- In an aircraft cockpit dialogue information system the pilot would expect the synthetic speech heard to be confident, clear and sometimes urgent, but never sympathetic or admonishing.

Global tones form the background tier for the pragmatic phonetic model. At this point global tones such as those expressing anger or happiness can be modelled. But more subtle attitudes, such as firmness and confidence, might well need to be modelled as the dialogue unfolds. In human dialogue there is a requirement to respond to pragmatic changes; the listener's perception varies as the context develops.

These changes can be characterized in another level superimposed on the global tone of voice, called local tone of voice.

2. Local tone of voice

Local tone of voice varies according to specific short term requirements during the unfolding dialogue. A speaker might be aware of a listener's changing levels of understanding while something is being explained and react accordingly with short term changes of style. As a specific example:

- A teacher needs to sound firm yet patient during the short term explanation of some point within a wider context.
- In the aircraft cockpit the computer's global firm and confident tone might be modified by encouraging and patient instructions if the pilot fails to understand an explanation or course of action.

3. The overall model

Tone of voice execution is modelled as a layered process. Execution begins with a neutral tone which might never be acoustically realized — an abstract representation of tone. This is the tone of 'neutral' phonology or the tone of a synthesis system implementing only a basic

language model, and is intended only for conveying plain messages.

Global tone is a specific long term modification of the abstract neutral tone. It is contextually determined by general pragmatic considerations deriving from the speaker.

Local tone comprizes specific short term *overlays* on global tone. It is contextually determined either by the changing nature of the semantics or pragmatics being communicated or by feedback concerning listener reaction. Local tone is superimposed on the global tone as the dialogue develops.

Both types of tone are generated by markers arising within the language model framework. Global markers are generated initially are only exceptionally updated, whereas local markers are repeatedly generated, updated or changed.

This framework is intended to relate observations of pragmatic effects in speech, to provide for a source for these effects (the pragmatic component in the language model), and will eventually set out the production and perceptual processes involved.

THE ACOUSTIC DATA

The acoustic data relating to pragmatic effects in speech is extremely difficult to obtain. It is not the intention of this paper to list acoustic correlates of particular pragmatic effects, but under this heading to account for some of the difficulties researchers face.

The biggest problem facing analysis of the acoustic signal is noise, that is, unwanted speech signal — not background noise against which the waveform is heard. The point here is that the variations imposed on the speech signal by pragmatic effects are buried in the natural variability associated with speech signals. Unfortunately it is not obvious which particular variation on any one occasion derives from the pragmatic marker — variability from many different sources is a basic characterisation of speech.

The problem is knowing what aspects of the variability are generated by pragmatically derived intentions and what aspects result from other sources. It was for this reason, for example, that data reduction was attempted using an artificial neural network paradigm [9] and a two-parameter (f0 and syllable durations)model to determine the associative relationship between pragmatic markers, abstract prosodic representations and an acoustic signal judged by listeners to evoke the required perceptual response. Despite the fact that neural networks are particularly good at tasks of this kind the results were disappointing: variations between speakers still became a confusing element in describing the acoustic signal.

Eskenazi [4] used a traditional technique of firstly selecting eight acoustic parameters (overall intensity, f0 maximum, dynamic range of f0, number of pauses, speaking rate, amount of phonological changes, F1/F2 shift and the amount of stop bursts) and then measuring them, but concluded that individual speakers expressed speech styles in different ways, and that not all parameters were equally used by all speakers. This phenomenon is also commented on by O'Shaughnessy [10], who emphasized that the mapping between physical acoustics and perceived prosody is not one to one.

Many researchers have tried hard to determine the acoustic correlates of pragmatic effects, and some have attempted to incorporate this information in their synthesis. So far, though, there has been little success in adequately unambiguously capturing subtle global effects like 'firmness' or providing sufficient contextual information to enable the automatic triggering of local effects.

CONCLUSION

The model most of us currently use assumes that the problem is to generalize acoustic cues from the waveform information. The listener is seen as responding to the cues. But if looked at from the listener's point of view he/she is supplying information in order to decode the signal. The interpretation of the signal will vary, but so will the acoustic cues responsible for triggering the percept.

If we wish to simulate the dialogue context either for practical purposes or to test the model we might well look more closely at modelling the listener, and how the listener may anticipate cues from knowledge of the dialogue context as it unfolds.

If the acoustic signal has within it cues for triggering an appropriate perceptual response to the pragmatics of the spoken utterance then that signal shows large variability and useful information is buried in variable noise.

Rather than model the process as heavily dependent on these cues, it is becoming necessary to shift the focus of the model away from the cues, leaving them as minimal and variable, and move toward a compensatorily oriented peceptual model. In traditional terms, we would assume a knowledge based system very heavily dependent on a full and accurate representation of the effects of various types of variability in the acoustic signal.

For the moment we cannot do this, but it is to be hoped that the adopting a model *framework* along the lines of what is suggested here might advance the situation. At the moment it is discouraging to measure acoustic data without some improvements to the framework within which that measurement is carried out.

REFERENCES

[1] Browman, C.P. and Goldstein, L. (1986), 'Towards an articulatory phonology', in C. Ewan and J. Anderson (eds.), *Phonology Yearbook 3*, Cambridge: Cambridge University Press, pp. 219–252.

[2] Tatham, M.A.A. (1986), 'Towards a cognitive phonetics', *Journal of Phonetics*, Vol. 12, pp. 37–47.

[3] Levinson, S.C. (1983), *Pragmatics*, Cambridge: Cambridge University Press.

[4] Eskenazi, M. (1992), 'Changing speech styles: strategies in read speech and casual and careful spontaneous speech', *Proceedings of the International Conference on Spoken Language Processing*, Banff, pp. 755–759.

[5] Bladon, A., Carlson, R., Granstrom, B., Hunnicutt, S. and Karlsson, I. (1987), 'A text-to-speech system for British English, and issues of dialect and style', *Proceedings of the European Conference on Speech Technology*, Edinburgh, pp. 55–58.

[6] Morton, K. (1993), 'Speech synthesis in dialogue systems', *Proceedings of Eurospeech '93*, Vol. 2, Berlin, pp. 905–908

[7] Granstrom, B. and Nord, L. (1992), 'Neglected dimensions in speech synthesis', *Speech Communication*, Vol. 11, pp. 347–356.

[8] Engstrand, O. (1992), 'Systemacity of phonetic variation in natural discourse', *Speech Communication*, Vol. 11, pp. 337–346.

[9] Morton, K. (1992), 'Pragmatic phonetics', in W.A. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing*, Vol. 2, London: JAI Press, pp. 17–53.

[10] O'Shaughnessy, D. (1987), *Speech Communication: Human and Machine*, Reading, Mass.: Addison-Wesley.